A Segmentation-free Approach for Keyword Search in Historical Typewritten Documents

B. Gatos, T. Konidaris, K. Ntzios, I. Pratikakis and S.J. Perantonis

Computational Intelligence Laboratory, Institute of Informatics and Telecomunications,
National Center for Scientific Research "Demokritos", GR-153 10 Agia Paraskevi, Athens, Greece
http://www.iit.demokritos.gr/cil, {bgat,tkonid,ntzios,ipratika,sper}@iit.demokritos.gr

Abstract

In this paper, we propose a novel segmentation-free approach for keyword search in historical typewritten documents combining image preprocessing, synthetic data creation, word spotting and user's feedback technologies. Our aim is to search for keywords typed by the user in a large collection of digitized typewritten historical documents. The proposed method is based on: (i) image preprocessing for image binarization and enhancement, noisy border and frame removal, orientation and skew correction; (ii) creation of synthetic image words from keywords typed by the user; (iii) word segmentation using dynamic parameters; (iv) efficient feature extraction for each image word and (v) a retrieval procedure that is optimized by user's feedback. Experimental results prove the efficiency of the proposed approach.

1. Introduction

Indexing historical typewritten documents is essential for quick and efficient content exploitation of the valuable historical collections. Our research focuses on keyword search in historical typewritten Greek documents (see Fig.1a) that date since the period of Renaissance and Enlightenment (1471-1821) and are among the first Greek typewritten historical documents. This work is developed under the framework of an integrated system that will manage and provide access to old Greek historical typewritten documents.

Traditional approaches to document indexing usually involve an OCR step [1]. In the literature two general approaches can be identified: the segmentation approach [2] and the global or segmentation-free approach [3]. The segmentation approach requires that each word has to be segmented into characters while the global approach treats each word as a single entity. Due to document degradations, it is not often feasible to get a correct segmentation of the typewritten historical documents into individual characters. Therefore, a segmentation-free approach must be employed. Such an approach is followed in [4] where line and word segmentation is used for word matching. In the segmentation-free approach there are works where word matching is based on the vertical bar patterns [5], as well as on weighted Hausdorff distance [6]. In the case of historical documents, Manmatha and Croft [7] presented a method for word spotting wherein matching was based on the comparison of entire words rather than individual characters. In this method, an off-line grouping of words in a historical document and the manual characterization of each group by the ASCII equivalence of the corresponding words are required. The volume of the processed material was limited to a few pages. This process can become very tedious for large collections of documents.

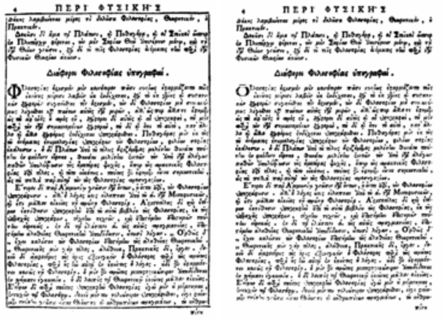

(a) (b)

Figure 1. A typical page from the historical typewritten Greek document collection. (a) Original image; (b) Resulting image after frame removal

To eliminate this tedious process, we propose a novel method for keyword search in historical typewritten documents which is based on: (i) image preprocessing; (ii) creation of synthetic image words from keywords typed by the user; (iii) word segmentation using dynamic parameters; (iv) efficient feature extraction for each image word and (v) a retrieval procedure that is supported by user's feedback. The synthetic keyword image is used to initialize the word spotting procedure and is matched against all detected words. The results are optimized

1520-5263/05 $20.00 © 2005 IEEE

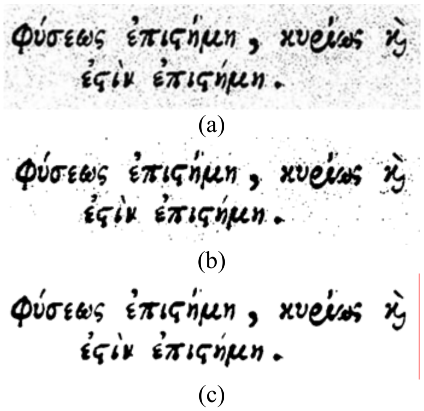
by the user's feedback. Combination of synthetic data creation and user's feedback leads to improved results in terms of precision and recall.

In the following sections, we present our workflow for keyword search in historical typewritten documents, as well as experimental results that demonstrate the efficiency of the proposed workflow.

2. Preprocessing

2.1. Image Binarization and Enhancement

Binarization is the starting step of most document image analysis systems and refers to the conversion of the gray-scale image to a binary image. Since historical document collections are most of the times of very low quality, an image enhancement stage is essential. The proposed scheme for image binarization and enhancement is based on the work of [8] and consists of five distinct steps: a pre-processing procedure using a low-pass Wiener filter, a rough estimation of foreground regions, a background surface calculation by interpolating neighboring background intensities, a thresholding by combining the calculated background surface with the original image and finally a post-processing step that improves the quality of text regions and preserves stroke connectivity. Fig. 2 illustrates image binarization and enhancement.



(a)

(b)

(c)

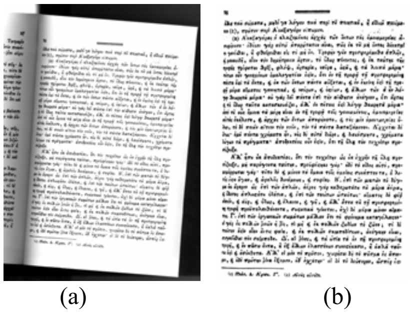
Figure 2. Image binarization and enhancement example. (a) Original gray scale image; (b) Resulting image after binarization; (c) Resulting image after image enhancement.

2.2. Orientation and skew correction

It is necessary to identify and correct the text orientation (portrait or landscape) and skew before proceeding to the word segmentation phase. Text orientation is determined by applying a horizontal/vertical smoothing, followed by a calculation procedure of vertical/horizontal black and white transitions [9]. For skew detection, we use a fast Hough transform approach based on the description of binary images using rectangular blocks [10].

2.3. Noisy Border removal

It is very frequent, the images resulting from scanning to be framed either by a solid or stripped black border (see Fig. 3a). In the preprocessing phase, we remove this border by employing a "flood-fill" based algorithm that moves from the outside noisy surrounding border towards the text region [11]. Additionally, image projections are used at the de-skewed image in order to remove noisy text from neighboring pages. Fig. 3 illustrates the application of our preprocessing scheme to a historical document suffering from skew and noisy border.



(a) (b)

Figure 3. Document preprocessing. (a) Original gray scale image; (b) Resulting image after preprocessing.

2.4. Average Character Height Estimation

The average character height estimation is required mainly for the smoothing process described in Section 3. For the calculation of the average character height we take the following steps:

STEP 1: We take a random pixel (x_A,y_A) that has at least one background pixel in its 4 connected neighborhood.

STEP 2: Starting from pixel (x_A,y_A), we follow the contour of the connected component that pixel (x,y) belongs to.

STEP 3: We repeat steps 1, 2 for all existing connected components until we have a maximum number of samples. During this process we calculate the histogram of the surrounding rectangles height at the corresponding connected components.

STEP 4: We compute the maximum value of the histogram which expresses the average character height.

2.5. Frame removal

To ease the segmentation process we remove potential frames around the text areas. The process of frame removal is based on the work of [12] and it is described in the following:

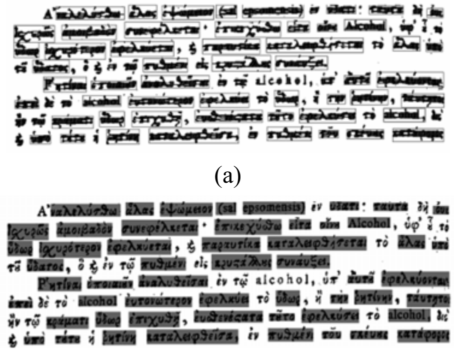
STEP 1: A sub-sampled version of the image is taken.
STEP 2: Two gray scale images F_H and F_V are extracted out of the sub-sampled image. In F_H every pixel has a value which corresponds to the length of the corresponding row of the sub-sampled image while in F_V every pixel has a value which corresponds to the length of the corresponding column.
STEP 3: The set of line segments that belong to a frame is characterized by an a priori known minimum length and maximum width. Images resulting from the above step are thresholded so that the only remaining pixels belonging to line segments whose length/width is larger/smaller than a threshold, respectively. The frame removal procedure is demonstrated in Fig. 1.

3. Segmentation

Our workflow involves the segmentation of the typewritten historical document images into words. This is accomplished with the use of the Run Length Smoothing Algorithm (RLSA) [13] by using dynamic parameters which depend on the average character height as it is described in section 2.4. RLSA examines the white runs existing in the horizontal and vertical directions. For each direction, white runs having length less than a threshold T_{max} are eliminated. In the proposed method, the horizontal length threshold is defined as half of the average character height while the vertical length threshold is set to a 10% value of the average character height. The application of RLSA results in a binary image where characters of the same word get connected to a single connected component (Fig 4a). In the sequel, a connected component analysis is applied using constraints which express the minimum expected word length (Fig 4b). This will enable us to reject stop-words and therefore eliminates undesired word segmentation.

4. Synthetic Data Creation

Synthetic data creation concerns the synthesis of the keyword images from their ASCII equivalences. A prerequisite is to have stored for all of the required characters their corresponding image template. During the manual character marking, an adjustment of the baseline for each character image template is supported in order to minimize alignment problems.



(a)

(b)

Figure 4. Segmentation process: (a) Resulting image after RLSA application; (b) Final word segmentation.

5. Word Retrieval

The word retrieval phase consists of two distinct steps; feature extraction and word matching. The segmented words are first set in a normalized bounding box preserving their aspect ratio.

5.1. Feature Extraction

The technique to carry out the word matching consists of extracting a set of features from the word images. Several features and methods have been proposed based on strokes, contour analysis etc. [14][15]. In our approach, we use two different types of features. The first one, which is based on [1], divides the word image into a set of zones and calculates the density of the character pixels in each zone. The second type of features is based on the work in [15]; in the proposed method we calculate the area that is formed from the projections of the upper and lower profile of the word.

5.2.1. Features based on zones. The image is divided into horizontal and vertical zones. In each zone, we calculate the density of the character pixels (see Fig. 5).

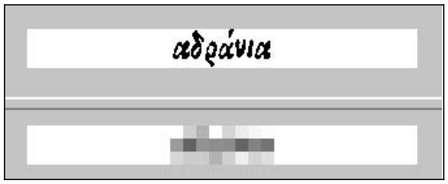


Figure 5. Feature extraction of a word image based on zones. The top frame shows the normalized word image while the bottom frame depicts the valuation of

the extracted features. Darker squares indicate higher density of character pixels.

5.2.2. Features based on word (upper/lower) profile projections. The word image is divided into two sections with respect to the horizontal line $y = y_t$ which passes through the center of mass (x_t, y_t) of the word image. Upper/lower word profiles are computed by recording, for each image column, the distance from the upper/lower boundary of the word image to the closest character pixel [15]. The word is divided into y_w vertical zones. For each upper/lower word profile we calculate the area in the upper and lower sections that correspond to the desired features (see Fig. 6).

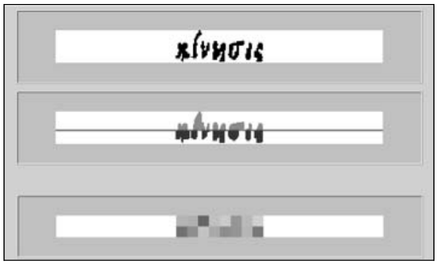


Figure 6. Top frame: The normalized word image; Middle frame: Upper and lower word profiles; Bottom frame: The extracted features. Darker squares indicate higher density of zone pixels.

5.3. Word Matching

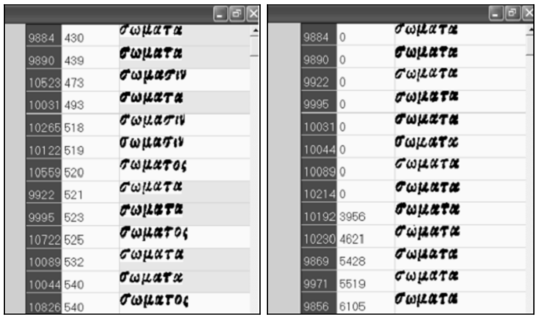
The process of word matching involves the comparison/matching of a synthetic keyword with all the segmented words found in all documents. Word matching is performed using the Manhattan Distance. Results are ranked and displayed accordingly (see Fig. 7a).

6. User's Feedback

User's feedback is an efficient mechanism for drastically improving the results of the matching process. The user selects the correct results from a list produced after the word matching process as described in Section 5.3. Then, a new matching process is initiated taking into account only those words that have been selected. The critical impact of the user's feedback in the word retrieval process lies upon the use of both real and correct data selected by the user. Furthermore, in our approach user interaction is supported by a simplified and user friendly human/machine interface that makes the word selection procedure an easy task. Fig. 7 illustrates the matching results before and after the user's feedback process.

7. Experimental Results

In order to evaluate the performance of the proposed method for keyword search in historical typewritten documents, we used the following methodology. We created a ground truth set by manually marking certain keywords on a subset of our document collection. The performance evaluation method used is based on counting the number of matches between the words detected by the algorithm and the marked words in the ground truth. For our experiments we used a sample of 50 document pages. The total number of words detected is 13,799. The task is to search for the Greek words W1 ("κίνησις") and W2 ("σώματα").



9884	430	σωματα
9890	439	σωματα
10523	473	σωματιν
10031	493	σωματα
10265	518	σωματιν
10122	519	σωματιν
10559	520	σωματος
9922	521	σωματα
9995	523	σωματα
10722	525	σωματος
10089	532	σωματα
10044	540	σωματα
10826	540	σωματος

(a)

9884	0	σωματα
9890	0	σωματα
9922	0	σωματα
9995	0	σωματα
10031	0	σωματα
10044	0	σωματα
10089	0	σωματα
10214	0	σωματα
10192	3956	σωματα
10230	4621	σώματα
9869	5428	σωματα
9971	5519	σωματα
9856	6105	σωματα

(b)

Figure 7. Searching for the Greek word "σωματα" with user's feedback: (a) Initial results list using synthetic data. Marked results indicate the user's selection of the correct words; (b) The final results of the word matching process.

Evaluation is performed using precision versus recall curves. Precision is the ratio of the number of relevant words to the number of retrieved words. Recall is the ratio of the number of relevant words to the number of total words marked on the images. We have used a variety of answer sets by a step of 10% of the total word instances in the dataset of the corresponding class.

The size of the normalized word images used is 300x30. In the case of features based on zones, the word image is divided into three horizontal and ten vertical zones. In the case of features based on word (upper/lower) profile projections, there are ten vertical zones and two horizontal zones. The total number of word instances is 60 for W1 and 37 for W2.

Fig. 8, 9 show the precision-recall rates for word images W1 and W2, respectively. In all cases when the user's feedback is applied, the precision/recall rates are improved. Combination of synthetic data creation and user's feedback leads to satisfactory results in terms of precision and recall. Furthermore, the use of features based on word (upper/lower) profile projections gives better results than using features based on zones.

8. Conclusions

This paper proposes a novel approach for keyword search in historical typewritten documents based on several distinct steps. By using a method of creating synthetic word images from their ASCII equivalences we achieved word search while avoiding the process of character recognition that occurs in traditional indexing systems based on OCR. The user is able to further improve the results by selecting the most relevant words from the list of the results and feed them back to the system. User's feedback enables the transition from synthetic queries to real data queries. This transition is fast and resources saving.

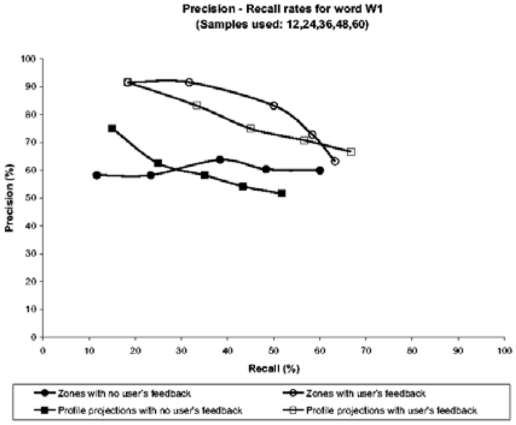


Figure 8. Precision - Recall rates for word image W1.

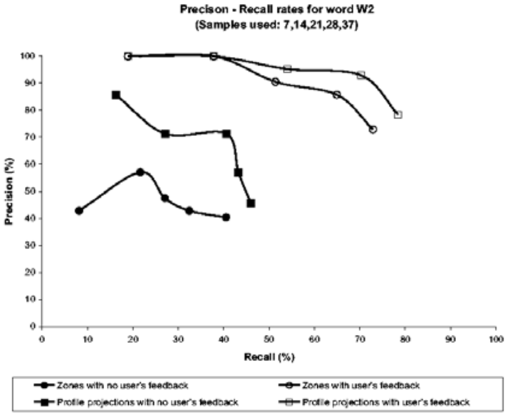


Figure 9. Precision - Recall rates for word image W2.

9. References

[1] M. Bokser, "Omnidocument technologies", Proc. of the IEEE, 80(7), 1992, pp. 1066-1078.

[2] B. Gatos, N. Papamarkos and C. Chamzas, "A binary tree based OCR technique for machine printed characters", Engineering Applications of Artificial Intelligence, 10(4), 1997, pp. 403-412.

[3] D. Guillevic and C.Y. Suen, "HMM word recognition engine", Proc. of the 4th Int. Conf. on Document Analysis and Recognition (ICDAR'97), 1997, pp. 544-547.

[4] A. Marcolino, V. Ramos, M. Ármalo and J.C. Pinto, "Line and Word matching in old documents", Proc. of the 5th IberoAmerican Sympsium on Pattern Recognition (SIARP'00), September 2000, pp. 123-125.

[5] H. Weihua, C.L. Tan, S.Y. Sung and Y. Xu, "Word shape recognition for image-based document retrieval", Proc. of the Int. Conf. on Image Processing, ICIP'2001, October 2001, pp. 8-11.

[6] Y. Lu, C. Tan, H. Weihua and L. Fan, "An approach to word image matching based on weighted Hausdorff distance", Proc. of the 6th Int. Conf. on Document Analysis and Recognition (ICDAR'01), September 2001, pp. 10-13.

[7] R. Manmatha and W.B. Croft, "A Draft of Word Spotting: Indexing Handwritten Manuscripts", Intelligent Multimedia Information Retrieval, MIT Press, Cambridge, MA, 1997, pp. 43-64.

[8] B. Gatos, I. Pratikakis and S.J. Perantonis, "An adaptive binarisation technique for low quality historical documents", IARP Workshop on Document Analysis Systems (DAS2004), Lecture Notes in Computer Science(3163), September 2004, pp. 102-113.

[9] P.Y. Yin, "Skew detection and block classification of printed documents", Image and Vision Computing 19, 2001, pp. 567-579

[10] S.J. Perantonis, B. Gatos and N. Papamarkos, "Block decomposition and segmentation for fast Hough transform evaluation", Pattern Recognition, vol. 32(5), 1999, pp. 811-824.

[11] B.T. Avila and R.D. Lins, "A new algorithm for removing noisy border from monochromatic documents", Proc. of the 2004 ACM Symposium on Applied Computing, 2004, pp. 1219-1225.

[12] B. Gatos, S.L. Mantzaris, S.J. Perantonis and A. Tsigris, "Automatic page analysis for the creation of a digital library from newspaper archives", International Journal on Digital Libraries (IJODL), vol. 3(1), 2000, pp. 77-84.

[13] R. Kasturi, S. Bow, W. El-Masri, J. Shah, J. Gattiker and U. Mokate, "A System for Interpretation of Line Drawings", IEEE Trans. Patt. Anal. Mach. Intell. 12, 1990, pp. 978-991.

[14] D. Doerman, "The detection of duplicates in document image databases", Proc. of the 4th Int. Conf. on Document Analysis and Recognition (ICDAR'97), 1997, pp. 314-318.

[15] T.M. Rath and R. Manmatha, "Features for word spotting in historical documents", Proc. of the 7th Int. Conf. on Document Analysis and Recognition (ICDAR'03), 2003, pp. 218-222.